JOURNAL OF MATHEMATICAL ANALYSIS AND APPLICATIONS 68, 548–556 (1979)

Connectedness Conditions used in Finite State Markov Decision Processes

L. C. THOMAS

Department of Decision Theory, University of Manchester, Manchester M13 9PL, England

Submitted by R. Bellman

This note considers the conditions that have been put on the set of transition matrices of finite state Markov Decision Processes in order to prove the existence of optimal policies, and the convergence of algorithms. It is shown that no two of the nine conditions considered are equivalent.

1. INTRODUCTION

The object of this note is to see how some of the conditions that have been required of transition matrices in Markov Decision Processes are related to one another. We concentrate on the case where the state and action space is finite, and on the average cost criterion.

Consider a Markov Decision Process where $\{1,..., N\}$ is the set of states and there is a finite set $K\{i\}$ of actions available in state i. Let $r_i(k)$ be the immediate reward of taking action k while in state i, while $p_{ij}(k)$ is the probability that the next state of the system will be state j, when action k is taken in state i.

Lef f be a decision rule that chooses an action $k(i) \in K(i)$ for each state i. Denote by F the class of all decision rules, and by $\mathscr{P}$ the corresponding sets of all possible transition matrices. This set has the closedness property, described in [1], namely that if P_1 , $P_2 \in \mathscr{P}$, then for any i, $1 \leqslant i \leqslant N$, the matrix obtained from P_1 by replacing the ith row of P_1 by ith row of P_2 belongs to $\mathscr{P}$.

A general policy π consists of choosing rules f_1 , f_2 ,..., f_n ,... at each step of the process, where f_i is the rule with i periods to go. Define $V^n(\pi)$, the reward in n periods using policy π, as the vector

$$V^n(\pi)_i = r_i(f_n) + \sum_{j=1}^{N} p_{ij}(f_n)\, V^{n-1}(\pi)_j$$

where $V^0(\pi)$ is the scrap value. This is chosen according to the conditions of the problem, but is usually taken to be zero. If π is $f, f,..., f,...$, then it is a stationary policy and we denote it by f_∞ . The aim is to find the policy π that

0022-247X/79/040548–09$02.00/0

maximises $V^n(\pi)$, for all components. One way of finding the optimal policy is to use a value iteration algorithm, by defining

$$v(n+1)_i = \max_{k \in K(i)} \left\{ r_i(k) + \sum_{j=1}^{N} p_{ij}(k)\, v(n)_j \right\}, \qquad 1 \leqslant i \leqslant N, \quad n = 0, 1, 2, \ldots$$

where $v(0)$ can be defined arbitrarily. Brown [4] showed that $\{v(n) - ng\}$ is bounded uniformly in n, permitting the interpretation of $g_i = \lim_{n\to\infty} v(n)_i/n$ as the maximal expected return per period, or gain, starting from state i. We only consider the case where g_i is independent of i. Another way of calculating the optimal gain, g, is the relative value algorithm in [13], where we chose one special state, m, and define

$$g_n = \max_{k \in K(m)} \left\{ r_m{}^k + \sum_{j=1}^{N} p_{mj}(k)\, w_{n-1}(j) \right\},$$

$$g_n + w_n(i) = \max_{k \in K(i)} \left\{ r_i{}^k + \sum_{j=1}^{N} p_{ij}(k)\, w_{n-1}(j) \right\}, \qquad i \neq m,$$

$$w_n(m) = 0,$$

having chosen $w_0(i)$ arbitrarily. Under certain conditions g_n converge to g.

Recall that the states of a transition matrix of a Markov chain can be divided into equivalence classes of communicating states. States i and j are in the same class if there is an n_1, n_2 such that $p_{ij}^{n_1} > 0$ and $p_{ji}^{n_2} > 0$. Different authors use the same term to define different properties of these classes, so we state our terminology. If all the states in a class are recurrent or persistent, the class is called ergodic, and if for a state i, the set $\{n \mid p_{ii}^n > 0\}$ has highest common factor of 1, the state is aperiodic. If a chain has only one equivalence class which is ergodic and aperiodic, it is called regular, so for some n $p_{ij}^n > 0$ for all i, j. We say the transition matrix is single ergodic aperiodic, S.E.A., if it has several communicating classes but only one ergodic one, and this is also aperiodic. However [10] calls these chains "completely ergodic."

We now list some of the conditions that have been put on the class in various papers on Markov Decision Theory, and consider the connections between them.

Condition 1 (C.1)

$\forall P \in \mathscr{P}$, P is regular.

This condition was used by Brown [4] to prove that $\lim_{n\to\infty}(v(n) - ng)$ exists where g is called the gain of the vector and was used by Derman and Strauch [6] to show that there is always a stationary optimal policy, though they immediately extend the proof to hold under condition C.7.

Condition 2 (C.2)

$\forall P \in \mathscr{P}$, P is S.E.A. with a common persistent state r, and $p_{rr} > 0$.

This was introduced by Anthonisse and Tijms [2] as an easy condition to check, that also implies condition C.3.

Condition 3 (C.3)

There is an r, $N \geqslant 1$ and $\alpha > 0$ such that $(P_1 \dots P_N)_{ir} \geqslant \alpha, \forall P_1, \dots, P_N \in \mathscr{P}$.

White [13] introduced this condition to prove the convergence of the relative value algorithm for the average cost Markov Decision Process. It is somewhat akin to the Doeblin-condition introduced in [8]. White also showed that this condition implies that $\lim_{n\to\infty} v(n) - ng$ exists for all $v(0)$ and that the convergence is geometric.

Condition 4 (C.4)

$\forall P \in \mathscr{P}$, P is S.E.A. with a common persistent state r.

This is a weakening on C.2. It does not require the strong aperiodic condition $P_{rr} > 0$.

Condition 5 (C.5)

There is a $N \geqslant 1$ such that for $n \geqslant N$, $P_1 P_2 \dots P_n$ is S.E.A. where $P_i \in \mathscr{P}$.

This is the asymptotic Markov condition dealt with by Wolfowitz [14]. Anthonisse and Tijms [1] showed that it is equivalent to the following two conditions.

Condition 5′ (C.5′)

There is an integer $N \geqslant 1$ such that for $k \geqslant N$ and any $P_i \in \mathscr{P}$, $1 \leqslant i \leqslant k$, the matrix $P_1 \dots P_k$ is scrambling i.e. any two rows of $P_1 P_2 \dots P_k$ have a positive entry in the same column.

Condition 5″ (C.5″)

There is an integer $N \geqslant 1$ such that for each $k \geqslant N$ and any $P_i \in \mathscr{P}$, $1 \leqslant i \leqslant k$, there is a state r, such that $(P_1 \dots P_k)_{ir} \geqslant \alpha > 0$.

Thus C.5″ is a weakening of C.3 in which the column with all positive entries will be different for different choices of $P_1, \dots, P_k$. It is an easy exercise to show that White's proof of the convergence of the relative value algorithm [13] still holds under this condition. It also ensures that there is an average cost for any policy, stationary or non-stationary, since [1] shows that $(P_1 \dots P_n)_{ij}$ converges to π_j, so that there is a limiting distribution.

Condition 6 (C.6)

For any states i, j there is a $P \in \mathscr{P}$, and $n \geqslant 1$ such that $p_{ij}^n > 0$. This was introduced by Bather [3] and implies that there is an optimal stationary policy

in the average cost case, with the gain being independent of the initial state. He then extended the result to convex decision spaces. Hordijk [9] also used this condition to prove the existence of an optimal stationary policy.

Condition 7 (C.7)

$\forall P \in \mathscr{P}$, P is S.E.A.

Howard [10] showed that this ensured a bounded solution to the optimality equation in the average cost case. Using the result that $\lim_{n\to\infty} V^n(f_\infty) = ng\,\underline{1} + \underline{v}$, the optimality equation becomes

$$g + v_i = \max_{k \in K(i)} \left\{ r_i(k) + \sum_{j=1}^{N} p_{ij}(k)\, v_j \right\}.$$

This condition ensures that a g and $\underline{v}$, which satisfy this equation, exist and are bounded. Derman and Strauch [6] also showed that this condition is enough for there to be a stationary policy which is optimal.

Condition 8 (C.8)

There is a state s and a finite number B such that for any stationary policy f, if $N = \inf\{n \geqslant 1 \mid X_n = s\}$, then

$$\mathrm{Exp}\{N \mid X_0 = i\} \leqslant B, \forall i$$

where X_i is the state of the system in the ith period. This condition of boundedness on first arrival at state s can be applied in the denumerable state case and ensures that the optimality equation for the average cost case has a bounded solution, see Derman [5], Derman and Veinott [7] and Hordijk [9].

Condition 9 (C.9)

The Markov decision process is simply connected, i.e. the state set consists of a single connected class C, along with a set of states T that are transient under all policies, where

(i) $i, j \in C \Rightarrow P_{ij}^n > 0$, some $P \in \mathscr{P}$

(ii) $i \in C, j \in T \Rightarrow P_{ij}^n = 0, \forall n, \forall P \in \mathscr{P}$

Platzman [11] shows that this condition is enough for the relative value algorithm to converge. It ensure that the average gain per step g is independent of the starting state.

Several of the above conditions imply that $\lim_{n\to\infty} v(n) - ng$ exists for all choices of $v(0)$. Schweitzer and Federgruen [12] proved that a necessary and sufficient condition for the existence of this limit is that there exists a randomised maximal gain policy whose transition matrix is aperiodic. It is trivial to see that condition C.7 implies this condition, and hence by the following theorem that conditions 1, 2, 3, 4 and 5 also imply it. However, since the Schweitzer-

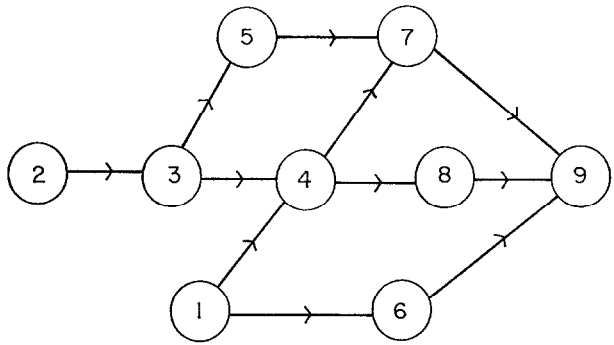
Federgruen condition depends on the rewards as well as the transition matrices, it cannot imply any of the above conditions. Neither do conditions 6, 8 or 9 imply it.

2. Relationship between the Conditions

No two of the above nine conditions are equivalent, and the relationships between them are given below.

THEOREM. C.2 $\Rightarrow$ C.3 $\Rightarrow$ C.4, C.5 $\Rightarrow$ C.7 $\Rightarrow$ C.9; C.1 $\Rightarrow$ C.6 $\Rightarrow$ C.9; C.1 $\Rightarrow$ C.4 $\Rightarrow$ C.8 $\Rightarrow$ C.9; *and none of the implications is an equivalence.*

Diagrammatically we have



Proof. C.2 $\Rightarrow$ C.3. The proof of this is given in Anthonisse and Tijms [2]. Define $S_0 = r$, $S_k = \{i \mid i \notin \bigcup_{n=0}^{k-1} S_n$ and for each $k \in K(i)$, there is a $j \in \bigcup_{n=0}^{k-1} S_n$ with $p_{ij}(k) > 0\}$, i.e. S_k are the states that you can get to r by every policy, for the first time after k steps. It is enough to show that if $\bigcup_{n=0}^{k-1} S_n \neq \{1,\ldots, n\}$, then $S_k \neq \{0\}$. Suppose the opposite, then there is a transition matrix P such that $p_{ij} = 0$, $\forall i \notin \bigcup_{n=0}^{k-1} S_n$, $j \in \bigcup_{n=0}^{k-1} S_n$. Thus $\{i \mid i \notin \bigcup_{n=0}^{k-1} S_n\}$ is a closed set which contradicts the fact that P is S.E.A. with persistent state s.

C.3 $\not\Rightarrow$ C.2. Take

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix} \quad \text{so} \quad P^2 = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix}$$

which satisfies White's condition but $P_{rr} = 0$, $r = 1, 2, 3$.

C.3 $\Rightarrow$ C.4. The implication is trivial. To show C.4 $\not\Rightarrow$ C.3, look at the case where

$$P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix},$$

so has persistent state $\{2\}$ and

$$P_2 = \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 1 & 0 & 0 \end{pmatrix}$$

with persistent states $\{1, 2, 3\}$, so satisfying C.4. However,

$$(P_1^2 P_2)^n = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} \quad \text{and} \quad P_1^n = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

so C.3 is not satisfied.

C.3 $\Rightarrow$ C.5. As was pointed out trivially, C.3 $\Rightarrow$ C.5″ which is equivalent to C.5. The counterexample showing C.5 $\not\Rightarrow$ C.3 is the same matrix as above where

$$P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \qquad P_2 = \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 1 & 0 & 0 \end{pmatrix}.$$

These do not satisfy C.3, but look at $P_i P_j$, P_i, $P_j \in \mathscr{P}$, i.e.,

$$P_2 P_1 = P_1^2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \qquad P_1 P_2 = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}, \qquad P_2^2 = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}.$$

Since they are all scrambling, and the product of any matrix with a scrambling matrix remains scrambling, P_1, P_2 satisfy C.5′ with $N = 2$.

C.4 $\Rightarrow$ C.7. Again the proof is trivial, while the counterexample showing C.7 $\not\Rightarrow$ C.4 has $\mathscr{P}$ consisting of the 8 possible matrices obtained from

Row 1: $(\frac{1}{2}, \frac{1}{2}, 0)$; $(\frac{1}{2}, 0, \frac{1}{2})$;
Row 2: $(\frac{1}{2}, \frac{1}{2}, 0)$; $(0, \frac{1}{2}, \frac{1}{2})$;
Row 3: $(\frac{1}{2}, 0, \frac{1}{2})$; $(0, \frac{1}{2}, \frac{1}{2})$.

These have no common persistent state, though all are S.E.A. In fact, this example also implies C.5 $\not\Rightarrow$ C.4, since all the matrices are scrambling.

C.5 $\Rightarrow$ C.7. Again the implication is trivial. To show C.7 $\not\Rightarrow$ C.5, let $\mathscr{P}$ consist of matrices constructed from

Row 1: $(0, \frac{1}{2}, \frac{1}{2}, 0)$; $(0, 0, \frac{1}{2}, \frac{1}{2})$;
Row 2: $(0, \frac{1}{2}, \frac{1}{2}, 0)$; $(\frac{1}{2}, \frac{1}{2}, 0, 0)$;
Row 3: $(\frac{1}{2}, 0, 0, \frac{1}{2})$; $(\frac{1}{2}, \frac{1}{2}, 0, 0)$;
Row 4: $(\frac{1}{2}, 0, 0, \frac{1}{2})$; $(0, 0, \frac{1}{2}, \frac{1}{2})$.

Let

$$P_1 = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \end{pmatrix} \quad \text{and} \quad P_2 = \begin{pmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

These are both S.E.A., but

$$P_1P_2 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

is not S.E.A. We have to check that any matrix of $\mathscr{P}$ is S.E.A., but since each row involves two entries, the only form of ergodic classes we can have is two of two states each, or one of four states. Since the first row involves two other states, it must be the latter. Aperiodicity is ensured by a positive diagonal term in the second row. Again this example also implies C.4 $\not\Rightarrow$ C.5, and C.1 $\not\Rightarrow$ C.5, since the matrices are regular.

C.7 $\Rightarrow$ C.9. In order that C.9 does not hold, the state set must be divided into two separate connected classes C_1, C_2 so that

(i) $i, j \in C_i \Rightarrow P_{ij}^n > 0$, some $P \in \mathscr{P}$

(ii) $i \in C_1, j \in C_2$ or $i \in C_2, j \in C_1 \Rightarrow P_{ij}^n = 0$, $\forall n$, $\forall P \in \mathscr{P}$.

Since under C.7 any $P \in \mathscr{P}$ is S.E.A., it has an ergodic class C which must be in one of the above connected classes of C.9, say C_1. Since all the other transient states i are such that for any state $j \in C_1$, $p_{ij}^n > 0$ for some n, none of them can be in C_2 as they cannot satisfy condition (ii). Thus C_2 will be empty and C.7 holds. The converse, C.9 $\Rightarrow$ C.7, is not true. Look at

$$\mathscr{P} = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right\}.$$

C.1 $\Rightarrow$ C.6 $\Rightarrow$ C.9. The implications again follow immediately from the definitions and the trivial example $\mathscr{P} = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ shows C.9 $\not\Rightarrow$ C.6, while

$$\mathscr{P} = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right\}$$

satisfies C.6 but not C.1. C.1 $\Rightarrow$ C.4 is again immediate, and

$$\mathscr{P} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

satisfies C.4 but not C.1.

C.4 $\Rightarrow$ C.8. This follows because any persistent state in a finite Markov chain is reached infinitely often, and so the time until first arrival is bounded. However, C.8 $\not\Rightarrow$ C.4 as the states in C.8 need not be aperiodic, i.e. $\mathscr{P} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. This example also shows that C.8 $\not\Rightarrow$ C.7, whereas the fact that C.7 $\not\Rightarrow$ C.8 is given by the counterexample for C.7 and C.4, since no state is recurrent for all the matrices in $\mathscr{P}$.

Lastly C.8 $\Rightarrow$ C.9. Recall that if C.9 is not to hold, there must be two connected classes C_1, C_2 such that for any $i \in C_1$, $j \in C_2$ then $P_{ij}^n = P_{ji}^n = 0$, $\forall n$, and $\forall P \in \mathscr{P}$. Let the persistent states of C.8 be in C_1, then since the expected first arrival from any other state j to s is bounded, we cannot have $P_{js}^n = 0$, $\forall n$. Thus there can be no states in C_2 and C.9 holds. Again the example

$$\mathscr{P} = \left\{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\right\}$$

shows C.9 $\not\Rightarrow$ C.8. In fact, this example also shows C.6 $\not\Rightarrow$ C.8.

To show there are no further equivalences between these conditions, it is sufficient to show C.6 $\not\Rightarrow$ C.7, and C.2 $\not\Rightarrow$ C.6. The first is given by the same example that shows C.6 $\not\Rightarrow$ C.8 above, while $\mathscr{P} = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ satisfies C.2 but not C.6.

References

1. J. M. Anthonisse and H. C. Tijms, Exponential convergence of products of stochastic matrices, *J. Math. Anal. Appl.* **59** (1977), 360–364.
2. J. M. Anthonisse and H. C. Tijms, "On White's Condition in Dynamic Programming," Report BW 46/75, Math. Centrum, Amsterdam, 1975.
3. J. Bather, Optimal decision procedures for finite Markov chains. II. Communicating systems, *Advances in Appl. Probability* **5** (1973), 521–540.
4. B. W. Brown, On the iterative method of dynamic programming on a finite space, discrete time Markov process, *Ann. Math. Statist.* **36** (1965), 1279–1285.
5. C. Derman, Denumerable state Markovian decision processes—average cost criterion, *Ann. Math. Statist.* **37** (1966), 1545–1553.
6. C. Derman and R. Strauch, A note on memoryless rules for controlling sequential control processes, *Ann. Math. Statist.* **37** (1966), 276–278.
7. C. Derman and A. Veinott, Jr., A solution to a countable system of equations arising in Markovian decision processes, *Ann. Math. Statist.* **38** (1967), 582–584.
8. W. Doeblin, Sur les propriétés asymptotiques de mouvements régis par certains types de chaînes simples, *Bull. Soc. Math. Roumaine* **39** (1937), 57–115.
9. A. Hordijk, "Dynamic Programming and Potential Theory," Mathematical Centre Tract No. 51, Math. Centrum, Amsterdam, 1974.
10. R. A. Howard, "Dynamic Programming and Markov Processes," Technology Press, Cambridge, Mass., 1960.
11. L. Platzman, Improved conditions for convergence in undiscounted Markov renewal programming, *Operations Res.* **25** (1977), 529–533.
12. P. J. Schweitzer and A. Federgruen, The asymptotic behaviour of undiscounted

value-iteration in Markov Decision Problems, *Math. Centre* Report BW 44/76; *Math. of O.R.*, in press.

13. D. J. WHITE, Dynamic programming, Markov chains and the method of successive approximations, *J. Math. Anal. Appl.* **6** (1963), 373–376.
14. J. WOLFOWITZ, Products of indecomposable, aperiodic, stochastic matrices, *Proc. Amer. Math. Soc.* **14** (1963), 733–737.